IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

A Survey on Network Intrusion System Attacks Classification Using Machine Learning Techniques

To cite this article: V. Deepa and N. Radha 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1022** 012036

View the article online for updates and enhancements.

A Survey on Network Intrusion System Attacks Classification Using Machine Learning Techniques

Mrs.V.Deepa[1], Dr.N.Radha[2]
[1]Assistant Professor, PSGR Krishnammal College for women, Coimbatore
[2]Associate Professor, PSGR Krishnammal College for women, Coimbatore

E-Mail:deepa@psgrkcw.ac.in[1],radha@psgrkcw.ac.in[2]

Abstract: Wireless Local Area Network (WLAN) security management is now being confronted by rapid expansion in wireless network errors, flaws and assaults. In recent times, as computers are used extensively through network and application creation on numerous platforms, attention is provided to network security. This definition includes security vulnerabilities in both complicated and costly operating programs. Intrusion is also seen as a method of breaching security, completeness and availability. Intrusion Detection System (IDS) is an essential method for the identification of network security vulnerabilities and abnormalities. A variety of significant work has been carried out on intrusion detection technologies often seen as premature not as a complete method for countering intrusion. It has also become a most challenging and priority tasks for security experts and network administrators. Hence, it cannot be replaced by more secure systems. Data mining used for IDS can effectively identify intrusion and the identified intrusion values are used to predict further intrusion in future. This paper presents a detailed review of literature about how data mining techniques were utilized for intrusion detection. First, intrusion detection on various benchmark and real-time datasets by data mining techniques are studied in detail. Then, comparative study is conducted with their merits and demerits for identifying the challenges in those techniques and then this paper is concluded with suggestions of solutions for enhancing the efficiency of intrusion detection in the network.

Keywords: *Network, Network security, Intrusion detection, Data mining and Machine Learning for IDS.*

1. Introduction

The network security is becoming more important aspects in computer networks because of rapidly increasing the usage in various applications and organizations. The existing network security tools like antiviruses, anti-spammers and firewalls used in many organizations can protect against network attacks. But, these tools cannot recognize new and complicated attacks.

Intrusion Detection (ID) [1] is a kind of security management system for networks and computers. In an ID system, information from different sources can be gathered and processed in a network or computer, detecting potential safety threats, all of which entail misuse and intrusions. A system which automatically monitor and analysis the event that occur in a computer network for detection of malicious activity is called Intrusion Detection System (IDS) [2]. Since the frequency of network security has significantly increased, the IDS have become a critical enhancement to most companies' security infrastructure. Intrusion detection allows the companies to their network from security breaches. Given the extent and complexity of current network security risks, it would not be a problem for security professionals to use intrusion detection, but rather to use the intrusion detection technologies and capabilities.

Intrusions are caused by attackers hacking databases, approved users trying to gain extra rights that they are not eligible for, and approved users who abuse their rights. In IDS, products look for attack signatures which specify suspicious or malicious intent [3]. Different techniques have been proposed to identify different types of intrusions, but there is no heuristic to confirm the accuracy of their results. Most of the conventional IDS depend on human analysts to distinguish intrusive and non-intrusive network traffic. However, it will take hours or days for detection or generation of new signature for an attack, which may involve infeasible to deal with rapid attacks. The network owners and operators are particularly concerned with information security when accessing the internet.
Because the internet world poses multiple threats of network attacks, various systems are built to prevent internet attacks.

Data Mining refers to the process of extracting models from large stores of data. The rapid development of data mining has made available a wide variety of algorithms, drawn from the fields of machine learning, deep learning .Several types of algorithms such as classification, link analysis and sequence analysis are particularly relevant to the intrusion detection work.

Data mining plays a very significant role in intrusion detection in the implementation of machine learning and provides approaches for the analysis of possible behaviors dependent on previous experiences. Hybrid association, clustering and classification are data mining techniques. Clustering is the division of data into groups dependent on results. The most widely used clustering technique is k-means clustering.

Classification and prediction are defined as the most common mining tool, which allows models to be retrieved, describes critical datasets and helps to forecast future patterns. Wahano [21] extends the data to be categorized as regular or abdominal in the IDS in metric-based classification. Classification split the data elements into one of the groups predefined. The performance of the classifier will be used to develop a system which will anticipate future patterns if the audit data captures sufficient regular and irregular behavior. The classification algorithm will forecast current unknown data by utilizing previously known data.

Nevertheless, the main problems presented by intrusion detection also tend to be demonstrated by major studies, which highlight the challenges of existing data mining equipment such as high False Alarm Ratio (FAR) and low Detection Rate (DR). However, in the same way significant researchers suggest that major challenges lie within intrusion detection and evidence demonstrates the difficulties in current data mining tools, such as high False Alarm Ratio (FAR), and low Detection Ratio (DR). In the current data mining tools, more improvements was suggested after concerns about the consistency of tools applied for mining audit data were posed.

Many Data mining techniques have been used in literature to detect the intrusions in the network effectively. Hence, the data mining techniques for intrusion detection in the literature is studied in detail this article. In addition, the advantages, performance efficiency and limitations are discussed. This detailed study provides a roadmap for finding new solutions to further improve the performance of IDS.

The content part of the paper is structured as follows: Section 2 presents the previous researches on intrusion detections using data mining technique in the literature. Section 3 lit out the performances of each technique by comparing in terms of merit, demerit and the performance efficiency, Section 4 talk about the findings in previous research works and Section 5 concludes an entire discussion and suggests new solutions for future enhancement.

2. Literature Survey

Chaïri et al. [4] proposed a sample selection method for IDS. It also solved the class imbalance problem in the dataset. The sample selection method generated a balanced dataset by paying more attention to those records located near the border line that enhanced the efficiency of classifier. Furthermore, the computational cost for sample selection was minimized by using clustering method. Initially in the clustering method, critical centers were found and then choose the samples from those critical clusters. Finally, the selected features were used in Multi-Layer Perceptron (MLP) for intrusion detection in the network.

Ambusaidi et al. [5] proposed a filter-based feature selection method to select optimal features which are used to classify the intrusions in the network. In the filter-based feature section method, theoretical analysis of mutual information was introduced which calculated the dependence between the features and output classes. An IDS called as Least Square Support Vector Machine based IDS (LSSVM-IDS) was used for intrusion detection. Filter-based feature selection method was used in this method for selecting most appropriate features. Flexible Mutual Information-based Feature Selection (FMIFS) was used for feature selection; this feature selection method did not require any parameter settings, because setting a suitable value of free parameters is a challenging task.

Varma et al. [6] proposed a fuzzy entropy-based heuristics for Ant Colony Optimization (ACO) method to search for global best subset of features for real time intrusion detection datasets. This method extracted various network traffic features with the range of discrete and continuous values. ACO utilized Fuzzy entropy as heuristic factor for selecting best feature from identified relevant features. The selected features increased the performance of intrusion detection classifiers. This method was best suitable for detecting various types of real time intrusion attacks.

Thaseenn & Kumar [7] proposed a paradigm of intrusion detection focused on the single class Support Vector Machine (SVM) chi square feature selection. Through extracting the variance for each feature, SVM kernel parameter has been optimized and the maximum feature variance of the attribute has been established. The kernel relied reverse on variance to boost the kernel parameters by a large variance. This intrusion detection model used a variance balancing methodology to refine the SVM parameters. It resulted in better classification accuracy for SVM.

Khammassi & Krichen [8] proposed a feature selection approach for IDS which produced optimal subset of features. Initially, the size of the dataset was reduced using re-sampling and then a wrapper approach was applied on the pre-processed dataset. The wrapper approach is the combination of genetic algorithm and logistic regression. From the wrapper approach, the best subset of features was selected and those features were used in NBTree, Random Forest (RF) and C4.5 classifiers to detect the intrusions in the network.

Raman et al. [9] proposed adaptive and robust IDS for feature selection and fine tuning of SVM parameters. In order to choose the most relevant features in the dataset and fine tune the SVM parameters for IDS, Hyper graph based Genetic Algorithm (HG-GA) was processed in the adaptive and robust IDS. The Hyper-Clique of Hyper graph was utilized in initial population that increased the convergence speed of IDS and also it helped to escape from local optima. Moreover, a weighted objective function was utilized in HG-GA to effectively balance the high detection rate and low false alarm rate.

Zhu et al. [10] proposed a scheme to choose the most relevant features for IDS. This scheme utilized several tailored searches and a specific strategy of control for population growth. This often distinguished traffic from regular or irregular behavior by abnormality. NSGA-III was utilized for the accomplishment of a properly executed feature subset, centered on this method. A modern niche survival technique was also introduced for an advanced multi-target optimization algorithm. Depending on the bias selection process, it chose the most relevant features. It selects the individual with the fewest functions and selects the entity with the greatest total weight of their goals in an acceptable selection process.

Aljawarneh et al. [11] proposed a hybrid model for anomaly-based IDS through feature selection analysis. A vote algorithm with information gain was applied in the collected data that combined the probability distribution of base learners to choose the most significant features. The selected features were used in different classifiers such as REPTree, AdaBoostM1, Meta Paging, Naïve Bayes and Random Tree for detection of intrusions in the network.

Roshan et al. [12] proposed an adaptive design of the IDS based on Extreme Learning Machine (ELM). A fast procedure was used in adaptive IDS to update the IDS based on the new patterns of data coming from existing and new attacks. The proposed IDS had the capability of obtaining input from a human security expert and being updated based on the input with the low computational cost. These approaches were used for cases where human experts requested for changing the cluster assignment existing data or a new class of data was available as the input. In such situation, this system updated the model without performing a full retraining process. However, this system achieved only an accepted rate of detection and false alarms.

Kabir et al. [13] proposed a novel technique based on sampling with Least Square Support Vector Machine (LS-SVM) for detection of intrusions in the network. This technique was comprised of two phases. In the first phase, the whole dataset was split into some pre-determined arbitrary subgroups. Then, the most discriminative features were chosen from these subgroups. The selected features resembled the whole dataset. After that, an optimum allocation scheme was explored according to the variability of observations within the subgroups. In the second phase of the proposed technique, LS-SVM was used to extract the samples for intrusion detection in a network.

Khan et al. [14] proposed a novel two-stage deep learning model for intrusion detection. Initially, the probability score value was calculated to categorize network traffics as abnormal and normal. The deep learner considered this probability score value as additional metric to take decision about intrusion state as normal or attack in the second stage. The probability score used in the second stage avoided over fitting problem of deep learner. This two-stage model capable to handle large amount of unlabeled data and able to learn features from large dataset effectively and automatically.

Zhang et al. [15] offered a Deep Belief Network (DBN) based intrusion detection model and DBN was optimized by improved genetic algorithm (GA). The number of number of hidden layer in DBN and the number of neurons in each layer was decided by improved GA. The results of the improved GA were used in DBN to detect the intrusions with compact structure and high detection rate.

Xiao et al. [16] proposed a Convolutional Neural Network (CNN) based network intrusion detection model with feature reduction method. The intrusion detection process was started with removing the redundant and irrelevant features through dimensionality reduction methods. After that, CNN was employed to extract the features of the reduced data. Moreover, the more effective information to identify intrusions was extracted using supervised learning.

Zhang et al. [17] proposed a deep hierarchical network for network intrusion detection. This network integrated LeNet-5 and LSTM while learning the spatial and temporal features of flow. The deep hierarchical network was trained at the same time instead of two training network through designing a reasonable network cascading method. The flow features have been examined to identify the network irregular flow.

Wei et al. [18] proposed an optimized Deep Belief Network (DBN) method for intrusion detection and classification. Initially, A Particle Swarm Optimization (PSO) was used in DBN to decide the learning factor and adaptive inertia weight. The initial optimization solution of PSO was optimized by fish warm behavior algorithm, the cluster of particle, foraging behaviors fish swarms decided the best initial PSO population. Once the best initial population solutions of PSO are generated, the global best solutions of PSO was searched by using genetic operators. Thus, PSO algorithm was optimized by fish swarm and genetic algorithms to obtain best optimal solutions for deciding the best DBN model for intrusion detection classification.

Yang et al. [19] proposed a combined network detection model based on deep learning model. A feature database was generated by feature mapping, one-hot encoding and normalization processing. A Deep Belief Network (DBN) with the multi-Restricted Boltzmann Machine (RBM) and Back-Propagation (BP) was constructed for intrusion detection. As an additional layer, the BP network layer was attached to the end of RBM. BP has been used to increase the weight of the multi-RBM. SVM has been used to prepare the system for intrusion detection.

Jiang et al. [20] proposed a network intrusion detection algorithm by integrating hybrid sampling and deep hierarchical network. A One-Side Selection (OSS) method was employed in the algorithm for removing the noise samples in the majority category. The minority samples were increased by applying Synthetic Minority Over-Sampling Technique (SMOTE). After solving the imbalance problem in the dataset, a Convolutional Neural Network (CNN) and Bi-directional Long

Short Term Memory (BiLSTM) were used to extract spatial and temporal features respectively those features were used for intrusion detection.

3. Comparative Analysis

A comparative analysis list out the merits and demerits of data mining techniques used for intrusion detection studied in the above section. Table 1 shows the efficiency and limitations of each method in simple representation. From the tabulated information, it can be easily concluded about the method which performs best among and the shot comings of each method. This comparison provides a road map to find new ideas for solving the shot comings of existing methods.

Table.1 Comparison of Different Data Mining Techniques based Intrusion Detection System

Ref No.	Methods	Merits	Demerits	Performance Metrics
[4]	Sample selection method, MLP	Selecting samples contribute positively in the improvement of the performance using sample selection method	Computation cost is high due to the multiple layers of computational units in MLP	Precision = 97.2%

[5]	Filter-based feature selection method, LSSVM-IDS, FMIFS	Low computational cost	Unbalanced sample distribution in the dataset.	Accuracy = 99.9%
				DR = 98.7%
				FPR = 0.28

[6]	Fuzzy entropy-based heuristics, ACO	Simple and faster way of detecting intrusions.	Time to convergence of ACO is uncertain	Average accuracy = 99.5%
[7]	Chi square feature selection, multi-class SVM	High classification accuracy	Proper selection of kernel function in SVM is more difficult	Accuracy=95.8%

[8]	Genetic algorithm, logistic regression, NBTree, RF, C4.5	Good detection rate	Failed to extract optimal subset of features that increase classification accuracy and decreases misclassified instances	Accuracy = 99.9%
				DR = 99.8%
				False Alarm Rate (FAR) = 0.105

[9]	HG-GA, SVM	High detection rate	SVM has some limitations like speed and size	Accuracy = 96.7% Detection rate = 97.1% False alarm rate = 0.83
[10]	NSGA-III	High classification accuracy	Bias threshold value greatly influence the detection rate	Accuracy = 99.2%
[11]	Hybrid model, REPTree, AdaBoostM1, Meta Pagging, Naïve Bayes and Random Tree	Minimize time complexity	It will not supported for fully distributed network	Accuracy = 99.2%
[12]	ELM	Low computational cost	Achieved only an accepted rate of detection and false alarms	Accuracy = 95.3% Detection rate = 0.67

[13]	LS-SVM	Can be used for both static and incremental data	High false alarm rate	Accuracy = 99.7% False alarm rate = 0.045

[14]	Novel two-stage deep learning model	High recognition rate	Unbalance class distribution affected the learning efficiency	Accuracy = 99.9%
				FAR = 0.00001%

[15]	Improved GA, DBN	High recognition rate, reduce complexity of network structure	Increase detection time	Accuracy = 99.4%
				Precision = 9.20%

[16]	Feature reduction method, CNN	Improves the classification performance	Not efficient for small number of attack categories	Accuracy = 97.1%
				DR = 0.96

[17]	LSTM, LeNet-5 neural network	High accuracy	To detect unknown types of attacks that have not been trained	Accuracy = 99.1%
				Precision = 99.3%

[18]	DBN, PSO	Shortens the average detection time	It greatly influence the training time	Accuracy = 83% False Positive Rate (FPR) = 0.77%

[19]	DBN, RBM, SVM	Good intrusion detection performance	It is not more efficient when the sample size of the network intrusion type is small	Accuracy = 97.4%
				Precision = 97.2%

[20]	Network intrusion detection algorithm, OSS, SMOTE, CNN, BiLSTM	Superior performance in terms of accuracy, precision and recall	Training time is high	Accuracy = 80.1%
				Precision = 80.8%

4. Findings

IDSs especially help to resist external attacks in the network. In the case that the standard Firewall can’t handle such tasks, IDSs can be used to prevent unauthorized network communication and operating system uses. **Cyber-attacks** will become sophisticated, so it is essential that protection technologies familiarize along with their threats.

Some methodologies of IDS are listed below.

Pattern change evasion: IDS depend on pattern matching concept to detect attacks. By making trivial adjust to the attack architecture, detection can be avoided.

Address spoofing/proxying: Attackers can opaque the source of the attack by using poorly secured or incorrectly configured proxy servers to leap an attack. If the source is spoofed and bounced by a server, it makes it very difficult to detect.

The comparative analyses of different data mining techniques for intrusion detection are employed mainly with two intrusion datasets NSL-KDD and UNSW-NB15 to evaluate the performance. Most of the research works concentrates on imbalance data distributions, convergence time, and distribution between normal and abnormal traffic, classification and detection rates. Based on performance metrics’ it could be concluded that the network intrusion algorithm combined with Deep Hierarchical network along with the deep learning algorithms such as OSS, SMOTE, BiLSTM[20] yields a superior performance in terms of accuracy, precision when compared to other algorithms.

5. Conclusion and Recommendations

In this paper, a survey on data mining techniques based IDS in the network are presented in detail. Also, merits and demerits of these techniques are discussed to suggest the future directions towards increase the performance of intrusion detection which effectively enhance the IDS. The results of comparative analysis proved that the network intrusion detection using deep hierarchical Network achieves better performance in terms of accuracy, precision and recall. However, the training time of network intrusion detection algorithm is high. Because the wireless network intrusion efficiency and accuracy are insufficiently measured in the above comparison, a deep learning-based model for network detection is proposed. The ability of deep learning in automatic feature extraction and feature selection reduces the difficulties in computing domain specific, hand-engineered features, and helps us to bypass the traditional feature selection phase. Deep learning (DL) is widely used in various fields and achieved good results [5].So in future, deep learning algorithms will be used to enhance the performance of network intrusion detection system by preventing the over fitting problem due to 0 elements, handling the issue of feature learning in a small number of attack categories, avoiding the misleading of DNN due to generation of adversarial input and finally to address the problem of unpredictability of Cyber-attacks.

References

[1] Mohit S D, Gayatri B K, Vrushali G M, Archana L G and Namrata R. B (2015). Using Artificial Neural Network Classification and Invention of Intrusion in Network Intrusion Detection System. *International Journal of Innovative Research in Computer and Communication Engineering*, ***3*****(2).**

[2] Zaman S, El-Abed M and Karray F (2013 January). Features selection approaches for intrusion detection systems based on evolution algorithms.

[3] Nazir A (2013). A Comparative Study of different Artificial Neural Networks based Intrusion Detection Systems. *International Journal of Scientific and Research Publications*, ***3*****(7), 1-15.**

[4] Chaïri I, Alaoui S, and Lyhyaoui A (2012, September). Intrusion detection based sample selection for imbalanced data distribution. In *Innovative Computing Technology (INTECH), 2012* (**pp. 259-264**). IEEE.

[5] Ambusaidi M A, He X Nanda P and Tan Z (2016). Building an intrusion detection system using a filter-based feature selection algorithm. *IEEE transactions on computers*, *65*(10), 2986-2998.

[6] Varma P R K, Kumari V and Kumar S S (2016). Feature selection using relative fuzzy entropy and ant colony optimization applied to real-time intrusion detection system. *Procedia Computer Science*, *85*, **503-510**.

[7] Thaseen I S and Kumar C A (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University-Computer and Information Sciences*, *29*(4), **462-472.**

[8] Khammassi C and Krichen S (2017). A GA-LR Wrapper Approach for Feature Selection in Network Intrusion Detection. *Computers & Security, 70,* **255-277**.

[9] Raman M G, Somu N, Kirthivasan K, Liscano R and Sriram V S (2017). An efficient intrusion detection system based on hyper graph-Genetic algorithm for parameter optimization and feature selection in support vector machine. *Knowledge-Based Systems*, ***134,*** **1-12.**

[10] Zhu Y, Liang J, Chen J and Ming Z (2017). An improved NSGA-III algorithm for feature selection used in intrusion detection. *Knowledge-Based Systems*, *116*, **74-85.**

[11] Aljawarneh S, Aldwairi M and Yassein M B (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*, *25*, **152-160.**

[12] Roshan S, Miche Y, Akusok A and Lendasse A (2018). Adaptive and online network intrusion detection system using clustering and extreme learning machines. *Journal of the Franklin Institute*, *355*(4), **1752-1779**.

[13] Kabir E, Hu J, Wang H and Zhuo G (2018). A novel statistical technique for intrusion detection systems. *Future Generation Computer Systems*, *79*, **303-318.**

[14] Khan F A, Gumaei A, Derhab A and Hussain A (2019). A novel two-stage deep learning model for efficient network intrusion detection. *IEEE Access*, *7*, **30373-30385.**

[15] Zhang Y, Li P and Wang X (2019). Intrusion detection for IoT based on improved genetic algorithm and deep belief network. *IEEE Access*, *7*, **31711-31722**.

[16] Xiao Y, Xing C, Zhang T and Zhao Z (2019). An intrusion detection model based on feature reduction and convolutional neural networks. *IEEE Access*, *7,* **42210-42219**.

[17] Zhang Y, Chen X, Jin L, Wang X and Guo D (2019). Network intrusion detection: Based on deep hierarchical network and original flow data. *IEEE Access*, *7*, **37004-37016.**

[18] Wei P, Li Y, Zhang Z, Hu T, Li Z and Liu D (2019). An optimization method for intrusion detection classification model based on deep belief network. *IEEE Access*, *7,* **87593-87605.**

[19] Yang H, Qin G and Ye L (2019). Combined Wireless Network Intrusion Detection Model Based on Deep Learning. *IEEE Access*, *7,* **82624-82632**.

[20] Jiang K, Wang W, Wang A and Wu H (2020). Network Intrusion Detection Combined Hybrid Sampling With Deep Hierarchical Network. *IEEE Access*, *8,* **32464-32476.**

[21] Wahono R S, “A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks,” *In Journal of Software Engineering*, **vol. 1,** April 2015, **pp. 1-16.**